Universal dynamical properties preclude standard clustering in a large class of biochemical data

Gomez, F ; Stoop, R-L ; Stoop, R

Abstract: Motivation: Clustering of chemical and biochemical data based on observed features is a central cognitive step in the analysis of chemical substances, in particular in combinatorial chemistry, or of complex biochemical reaction networks. Often, for reasons unknown to the researcher, this step produces disappointing results. Once the sources of the problem are known, improved clustering methods might revitalize the statistical approach of compound and reaction search and analysis. Here, we present a generic mechanism that may be at the origin of many clustering difficulties. Results: The variety of dynamical behaviors that can be exhibited by complex biochemical reactions on variation of the system parameters are fundamental system fingerprints. In parameter space, shrimp-like or swallow-tail structures separate parameter sets that lead to stable periodic dynamical behavior from those leading to irregular behavior. We work out the genericity of this phenomenon and demonstrate novel examples for their occurrence in realistic models of biophysics. Although we elucidate the phenomenon by considering the emergence of periodicity in dependence on system parameters in a low-dimensional parameter space, the conclusions from our simple setting are shown to continue to be valid for features in a higher-dimensional feature space, as long as the feature-generating mechanism is not too extreme and the dimension of this space is not too high compared with the amount of available data.

DOI: https://doi.org/10.1093/bioinformatics/btu332

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2486–2493
doi:10.1093/bioinformatics/btu332

Databases and ontologies

Advance Access publication May 10, 2014

Universal dynamical properties preclude standard clustering in a large class of biochemical data

Florian Gomez[1], Ralph L. Stoop[2] and Ruedi Stoop[1,*]

[1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, 8057 Zurich, Switzerland and [2]Institute of Physics, University of Basel, 4056 Basel, Switzerland

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Clustering of chemical and biochemical data based on observed features is a central cognitive step in the analysis of chemical substances, in particular in combinatorial chemistry, or of complex biochemical reaction networks. Often, for reasons unknown to the researcher, this step produces disappointing results. Once the sources of the problem are known, improved clustering methods might revitalize the statistical approach of compound and reaction search and analysis. Here, we present a generic mechanism that may be at the origin of many clustering difficulties.

Results: The variety of dynamical behaviors that can be exhibited by complex biochemical reactions on variation of the system parameters are fundamental system fingerprints. In parameter space, shrimp-like or swallow-tail structures separate parameter sets that lead to stable periodic dynamical behavior from those leading to irregular behavior. We work out the genericity of this phenomenon and demonstrate novel examples for their occurrence in realistic models of biophysics. Although we elucidate the phenomenon by considering the emergence of periodicity in dependence on system parameters in a low-dimensional parameter space, the conclusions from our simple setting are shown to continue to be valid for features in a higher-dimensional feature space, as long as the feature-generating mechanism is not too extreme and the dimension of this space is not too high compared with the amount of available data.

Availability and implementation: For online versions of super-paramagnetic clustering see http://stoop.ini.uzh.ch/research/clustering.

Contact: ruedi@ini.phys.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 13, 2014; revised on April 16, 2014; accepted on May 7, 2014

1 INTRODUCTION

Biological organisms are able to fabricate intricate machineries from the molecular scale up to the macroscopic scale, without the obvious need to store and to explicitly handle the corresponding information. Synthetic biology, molecular programming and nucleic acid nanotechnology have thus become an experimental playground for the search for systems that carry out human-defined molecular programs, to input, output and manipulate molecular structures (Ando *et al.*, 2011; Silva and McClenaghan, 2004). For chemistry to become the next information technology substrate, improved tools for designing, simulating and analyzing complex molecular circuits and systems are necessary. On the DNA nanotechnology model system, corresponding knowledge is presently quickly growing and the area of alternative computing paradigms starting to take shape. From a physics point of view, biological and physical processes start to converge, so that to describe biochemical computation, concepts from physics can be borrowed and applied (see e.g. Brackley *et al.*, 2010; Ellner and Guckenheimer, 2006; Furusawa and Kaneko, 2012).

As will be exhibited below (Figs 1 and 2), most real-world systems exhibit a non-trivial behavior of some observables in time. Many such processes exhibit periodicity (the circadian rhythm, the cell cycle, reproduction), which therefore has often been regarded as a key expression of the essential mechanisms of life. Conversely, irregular behavior is often related to abnormal stimuli or to a defect or disorder of the generating mechanism (the cortex, however, provides an example that shows that this does not necessarily have to be the case; see Stoop *et al.*, 2000). Modern methods of measurements and modeling have now provided techniques that permit the observation of dynamical aspects of processes, which in the past, because of a lack of such technology, were described as steady state. Genetic expression processes are an example thereof (Romano *et al.*, 2009). Recently, it has been possible to measure down to single-cell expression, which revealed different kinds of rhythmic to irregular expression patterns (Raj and van Oudenaarden, 2008; Spiller *et al.*, 2010; Suter *et al.*, 2011). In our study, we will put forward a generic model that demonstrates that regular and 'stochastic' expression may result from the same non-linear system and that the transition among these states may require small parameter changes only. In Section 5, we will exhibit how more general gene expression complexity may emerge from the generic model.

A particular well-known example of regular behavior is the circadian clock. Decades ago, the circadian rhythm was believed to be singularly implemented by means of a central clockwork or pattern generator. It was, however, discovered that the mutation of a single allele of a single locus (called the 'per' gene) triggers *Drosophila* mutants with different circadian rhythms (Konopka and Benzer, 1971). Many of the genes and proteins involved in this process have been evidenced in mammals as well, where the circadian clock arises from the temporally regulated activity of protein–gene pairs (Tei *et al.*, 1997; Ueda *et al.*, 2005). A variety of tissues and cells containing functional autonomous clocks are

*To whom correspondence should be addressed.

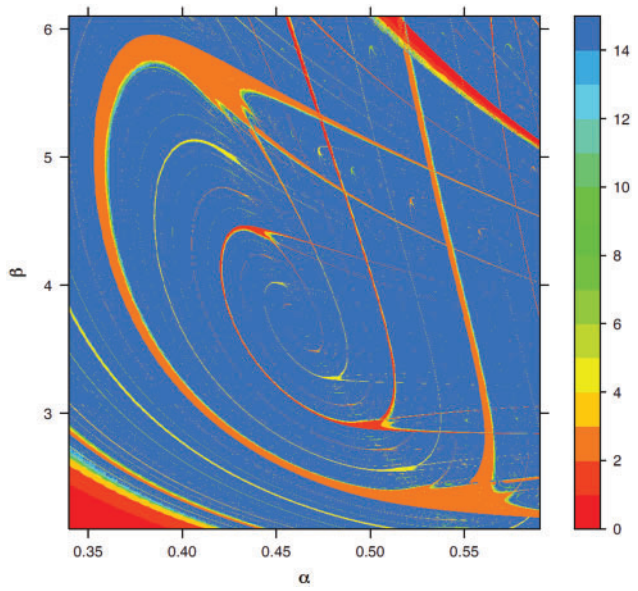

Fig. 1. Shrimps of the Inaba–Nishio electronic circuit: color-coded parameter space areas of fixed periodicities p_i. Parameters: α and β. In (Stoop *et al.*, 2010), it was shown that the dynamical behaviors of the hardware-built circuits (each pixel corresponds to a particular realizable circuit) follow exactly the predicted structure

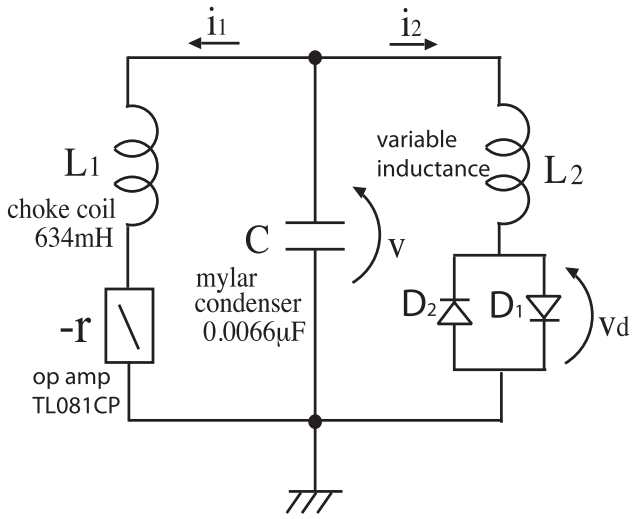

Fig. 2. Non-linear Inaba–Nishio circuit diagram cf. Stoop *et al.* (2010)

able to maintain their oscillation when placed *in vitro*, removed from any external cues or signals originating from environmental clocks. This naturally leads to the hypothesis that the circadian rhythm, and more generally periodic dynamics, is dispersed in any living system, with potentially every cell containing a functional clock. From the genetic base, the circadian rhythm is implemented by means of a gene expression network hosting a delayed feedback loop that causes the transcripts to oscillate with an approximate period of 24 h. A recently found period doubling in the mouse (Erzberger *et al.*, 2013) and oscillation bifurcations in the rat (Granada *et al.*, 2011) hint at the non-linear nature of this effect. Non-linearity is also at the heart of the phenomenon that we describe below. The aim of this article is to demonstrate how the distribution of regular and irregular systems is governed by a system's non-linearity in a generic manner and that from this, consequences for the system biology analysis of data may emerge. The two instances of system biological analysis that we discuss at the end of this contribution from this point of view are clustering on a system level and parameter inference. For biological systems, the identification of the areas in parameter space responsible for periodic behavior is therefore an important task.

2 HOW SYSTEMS ARE DISTRIBUTED IN PARAMETER SPACE: A COMMON PITFALL

Areas in a parameter space that exhibit in some sense common features are usually determined by a clustering process. In a one-dimension parameter space, the natural picture of systems displaying a certain periodicity would be given by parameters that are distributed across an interval, possibly according to a Gaussian (Murua *et al.*, 2008; Yeung *et al.*, 2001), where, for non-linear systems, these intervals would generally be finite. In higher-dimensional parameter spaces, following this reasoning, a Cartesian product of such intervals (i.e. in dimension two, a square or a circle, depending on the topology or distance function chosen) will be expected to guarantee the emergence of periodicity. This conclusion, which is the basis of many bioinformatics approaches (e.g. the popular k-means clustering), is wrong. For real-world systems, the generic parameter space domain for periodicity is a 'swallow-tail' or a 'shrimp'-like domain. Its convex–concave form was predicted by Shilnikov's theory (Shilnikov, 1965, 1967; Shilnikov *et al.*, 1998, 2001), and it was later discussed in more details by Gaspard, Kapral and Nicolis (Gaspard *et al.*, 1984). The emergence of shrimps has been evidenced in a number of models of physical systems. Most prominent examples are a model of the Inaba–Nishio simple resistive electronic circuit (Bonatto and Gallas, 2008; Nishio *et al.*, 1990; Stoop *et al.*, 2010) and a simple model of a CO_2 laser (Bonatto *et al.*, 2005). Although shrimps are easily detected in simulations, the experimental verification is more demanding, as because of their complex boundaries, a high experimental resolution is needed to pin them down. One of the first—at that time somewhat tentative example of an experimental shrimp—was provided for Chua's circuit (Baptista, 1996; Baptista *et al.*, 2003). Efforts focusing on the experimental verification of shrimps have continued ever since (Cardoso *et al.*, 2009; Maranhao *et al.*, 2008). To highlight that shrimps can be observed in real systems, we focus on the Inaba–Nishio resistive circuit. The Inaba–Nishio circuit contains a linear negative resistance ('$-r$'), a capacitance (C), two coils (L_1, L_2) and a non-linear resistance introduced either by two diodes (D_1, D_2, *'symmetric* circuit') or by one diode only (D_1, *'asymmetric* circuit'); see Figure 2. Depending upon parameter α (coding for a combination of the properties of resistance r, capacitance C and coil L_1) and β (coding for the properties of both coils L_1, L_2), the behavior of the system is characterized by a spiral of shrimps (Fig. 1; Stoop *et al.*, 2010).

3 EMERGENCE OF SHRIMPS IN PARAMETER SPACE

How is this multitude of scaled versions of the same shrimp template generated? In the case of smooth systems, shrimps are the result of the interaction of two or more largely independent parameters in creating points with a full set of zero partial derivatives. From this observation, the shrimps phenomenon can be explained in a simple way, for flows [the Rössler system (Gaspard *et al.*, 1984)] and for maps [the dissipative Hénon map (Hénon, 1976) in (Gallas, 1993, 1995)]. For simplicity of argument, we will consider the discrete formulation and follow the exposition given in Stoop *et al.*, 2010, 2012. Note that the dissipative Hénon map is the paradigmatic 2D discrete map accounting for the universality properties of dissipative non-linear systems. The Hénon map can be written in its standard form as (Kuznetsov, 2004) $f_h : \{x, y\} \to \{c - dy - x^2, x\}$. After cycling through the coordinates by means of two iterations, the 2D system can be condensed into the approximative 1D map

$$f: \; x \to b - (a - x^2)^2,$$

which incorporates the two parameters a, b for the offset and the leading term non-linearity in one equation.

Stable k-periodic islands arise whenever

$$x_k = f^k(x_k), \quad |m_k| = |f^{k'}(x_k)| < 1 \tag{1}$$

holds, where f^k denotes the k-fold iterated map f, and the prime $'$ denotes the derivative with respect to x. A superstable locus requires that $m_k = 0$. More explicitly, we have

$$f^{k'}(x_k) = \prod_{i=1}^{k} 4x_i \prod_{i=1}^{k} (a - x_i^2). \tag{2}$$

This implies that all k-superstable solutions need to pass either through $x_k = 0$ or $x_k = \pm\sqrt{a}$. For the case $x_k = 0$, for $k = 1$, we obtain from $b - (a - x^2)^2 = x$ the relation $a = \pm\sqrt{b}$. For the case $x_k = \pm\sqrt{a}$, we obtain $b = \pm\sqrt{a}$. By differentiability of f in the parameters a, b, this defines two parabolas in parameter space, which define the four legs of the main $k = 1$-shrimp; see Figure 3. The two parabolas intersect at points $\{0, 0\}$ and $\{1, 1\}$, giving rise to the 'head' and the 'navel' of the shrimp, respectively, manifested by two distinguished doubly superstable systems. Once more, by means of differentiability, an area around these lines is identified, within which

$$|f^{k'}(x_k)| = |\prod_{i=1}^{k} 4x_i \prod_{i=1}^{k} (a - x_i^2)| \leq 1 \tag{3}$$

is fulfilled, for the sequence of points x_i visited. For $k = 1$, we obtain from the fixed-point and the derivative conditions the pair of equations valid for the asymptotic behavior at the head of the shrimp

$$\{f^{k=1}(x) - x = 0, \; f^{k=1'}(x) - 4a\,x + 4x^3 = 0\}. \tag{4}$$

Using $\mu := f^{k=1'}(x)$ and elimination of the explicit phase-space variable x, we obtain

$$\mu^4 - 12\mu^3 + (48 - 32ab)\mu^2 + 64(ab - 1)\mu - 256(a - b^2)(a^2 - b) = 0.$$

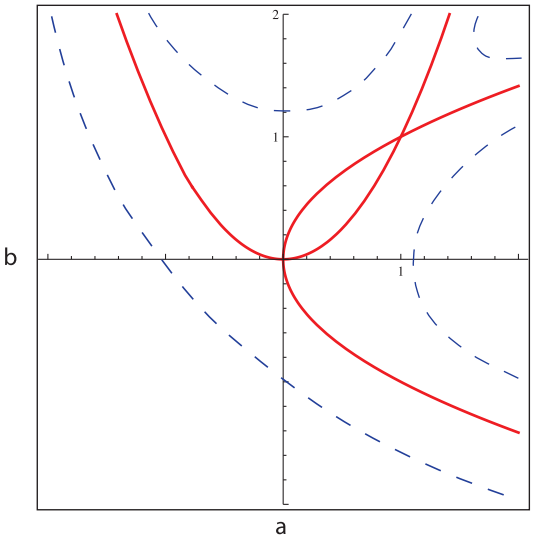

Fig. 3. Basic shrimps structure: two intersecting parabolas of superstability (red full lines), extending until the derivative of the solutions exceeds 1 in absolute value (non-generically located dashed lines), where tangent or period-doubling bifurcations occur. In addition, where lines cross, we deal with non-ergodicity (Stoop *et al.*, 2012). Secondary, non-generic, system properties can complicate this fundamental structure

From this equation, upon letting $\mu = 0$, the above-identified two parabolas

$$a = \pm\sqrt{b}, \; b = \pm\sqrt{a} \tag{5}$$

emerge. Restricting μ to values $|\mu| \leq 1$, we can identify the area of the period $k = 1$ shrimp in the $\{a, b\}$-parameter space. This area is bounded by period-doubling bifurcations ($\mu = -1$) and tangent bifurcations ($\mu = 1$) on opposite sides from the loci of superstability.

By representing Feigenbaum universality in higher-dimensional parameter space, the emergence of shrimp-like structures is thus a universal non-linear phenomenon, i.e. it must be expected to occur in any non-linear dynamical system. The placement of the copies is, however, determined by the specific system's properties (Stoop *et al.*, 2012).

4 BIOLOGICAL MANIFESTATIONS OF SHRIMPS

To what extent such structures emerge in biological systems has mostly remained unexplored. This is a non-trivial question because vast areas in parameter space may not be occupied by typical real-world biological systems and processes. Here we focus on two domains where the dependence of the dynamics on system parameters is of special interest: Biochemical reactions and neural systems.

Biochemical systems: For the field of biochemical reactions, we focus on an enzymic reaction, noting that periodic behavior is not exclusive to enzymic processes. We consider the celebrated Goldbeter reaction (Decroly and Goldbeter, 1982), for which corresponding experimental evidence is available (De la Fuente, 1999; Markus *et al.*, 1985). Enzymic periodicities are best described at the molecular level. On this level, the

Goldbeter reaction can be represented as shown in Figure 4: substrate S is injected at constant rate v and runs through a sequence of enzymic reactions comprising two positive feedback loops coupled in series. S is transformed by catalyzation by an enzyme E_1, which is activated by its product P_1. A second enzyme E_2 uses P_1 as substrate and is activated by its product P_2. k_s is the first-order rate constant for the removal of P_2. The two steps are necessary to generate, along with periodicity, chaotic behavior.

The metabolite concentrations can be described by the following three ordinary differential equations:

$$\frac{d\alpha}{dt} = \frac{v}{K_{m1}} - \sigma_1\phi,$$

$$\frac{d\beta}{dt} = q_1\sigma_1\phi - \sigma_2\eta,$$

$$\frac{d\gamma}{dt} = q_2\sigma_2\eta - k_s\gamma,$$

with

$$\phi = \alpha(1+\alpha)(1+\beta)^2/[L_1 + (1+\alpha)^2(1+\beta)^2],$$

$$\eta = \beta(1+d\beta)(1+\gamma)^2/[L_2 + (1+d\beta)^2(1+\gamma)^2].$$

α, β and γ denote the concentrations of S, P_1 and P_2 divided, respectively, by K_{m1}, K_{P_1} and K_{P_2}. K_{m1} is the Michaelis constant of E_1 for the substrate S, K_{p1} the dissociation constant of P_1 for E_1 and K_{P_2} is the dissociation constant of P_2 for E_2. v denotes the constant input of substrate, and σ_1 and σ_2 are the maximum activities of E_1 and E_2 divided by their Michaelis constants K_{m1} and K_{m2} (K_{m2} is the Michaelis constant of E_2 for its substrate P_1), respectively. L_1 and L_2 are the allosteric constants of E_1 and E_2, respectively. Finally, $q_1 = K_{m1}/K_{P_1}$, $q_2 = K_{P_1}/K_{P_2}$ and $d = K_{P_1}/K_{m2}$.

The two reaction steps are required to provide the system with the ability to produce irregular chaotic solutions. Although Decroly and Goldbeter (1982) considered changes of k_s and v, and reported no shrimps, we investigate here the behavior obtained by changing σ_1 and σ_2, for which we observe an abundant emergence of shrimps (Fig. 5). Clearly, we find shrimp-like structures with stable periodic oscillations, starting from period 4 (dark gray) to 8 (green) to 16 (yellow) to 32 (ocher). Domains of chaotic behavior are in white.

Neural systems: Neuron models share many structural properties of enzymatic reactions; the occurrence of shrimps in neuron models is therefore only surprising in light of the fact that so far, their existence has not been reported [discounting structures that vaguely resemble half-cuts of shrimps (Gallas, 2010)]. In a number of cases, a too low dimensionality of the model prevents chaotic behavior from occurring, whereas for other models, the huge number of coupled equations and parameters involved may be prohibitive for such a research (e.g. for Hodgkin-Huxley-like equations). The phenomenological neuron model elaborated by Rulkov (Rulkov, 2002; Shilnikov and Rulkov, 2003) does not suffer from these limitations and has been repeatedly shown to accurately describe the dynamics of biological neurons (Martignoli *et al.*, 2013; Nowotny *et al.*, 2005; Rulkov *et al.*, 2004; Tainaka *et al.*, 2006;) because of a versatility based on minimal modeling. The equations of this model are based on two parameters α and σ,

$$x_{n+1} = f(x_n, y_n), \tag{6}$$

$$y_{n+1} = y_n - \mu(x_n + 1) + \mu\sigma, \tag{7}$$

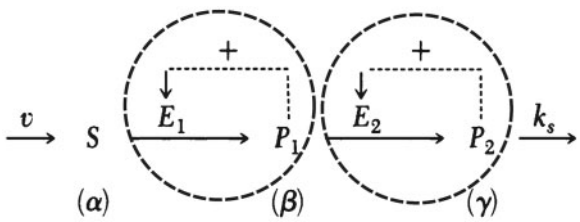

Fig. 4. Goldbeter's two-step biochemical feedback loop reaction process

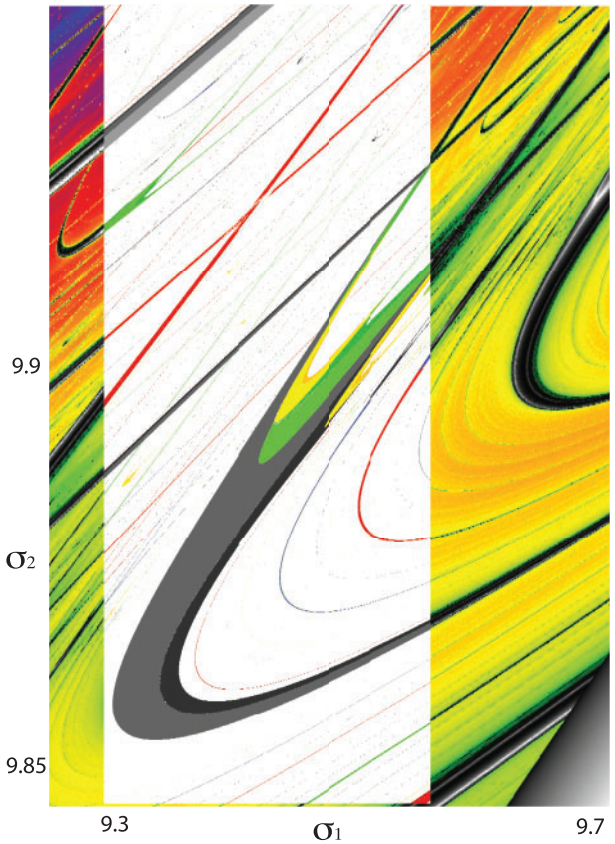

Fig. 5. Shrimp-like structure in the biochemical system described in Decroly and Goldbeter (1982). The parameter space is $\sigma_1 = 9.25 - 9.7$ and $\sigma_2 = 9.82 - 9.96$. Constants: $v = 0.45$, $k_s = 2.01$, $q_1 = 50$, $q_2 = 0.02$, $K_{m1} = 1.0$, $L_1 = 5 \times 10^8$, $L_2 = 100$, $d = 0$. Background: characterization of the parameter space properties in terms of the largest non-zero Lyapunov exponent color coded as shown in Figure 6. In the central part, we superimpose the observed periodicity of solutions. Chaotic motion generates infinite periodicity (white dots). Non-white dots correspond to finite stable periodicities. In the center of the figure, a period-doubling route to chaos is exhibited (gray → green → yellow → ocher). In the original work of Decroly and Goldbeter (1982), $\sigma_1 = \sigma_2 = 10$ were held constant

where μ is set to $\mu = 0.001$, and $f(x,y)$ is given by

$$f(x,y) = \begin{cases} \alpha/(1-x)+y, & x \le 0 \\ \alpha+y, & 0 < x < \alpha+y \\ -1, & x \ge \alpha+y, \end{cases} \quad (8)$$

where the first variable codes, in loose terms, for the inner (spiking) state of the neuron, and the second variable codes for a slower background state on which this dwells. The parameters α and σ that we will focus on encode (again in loose terms) the non-linearity and the driving current of the neuron, respectively. It is easily verified that regular, periodically firing behavior occurs on shrimp-like domains of the parameter plane (Fig. 6).

5 DIFFICULTIES FOR CLUSTERING AND SOLUTION

From these examples, it is evident is that in non-linear biological systems, steady-state behavior will often be the exception rather than the rule. Moreover, the proposition emerges that many of the processes that are currently declared stochastic may be chaotic. Biological systems may exploit both behaviors, preferentially even in symbiosis: A large number of small chaotic or stochastic inputs to a neuron, e.g. will generate an optimally stable input current that will force the neuron to fire regularly, generally on a limit cycle. The closer the generated response is to stochasticity, the better an ensemble of such systems provides a reliable constant driving current to the neuron, leading to a stable firing pattern of periodicity one. From the interaction among such oscillators, more complicated periodic patterns emerge, the periodicities of which are organized along Arnol'd tongues (Martignoli and Stoop, 2008; Stoop *et al.*, 2000). In the context of the circadian rhythm, the observation of locking on an Arnol'd tongue has recently been reported, along with period doubling (Erzberger *et al.*, 2013). Both are manifestations of non-linearity, within or among individual entities. The former effect occurs when the coupling is relatively small; the latter effect occurs and dominates when the coupling is 'larger'. The manifestation in both cases is the emergence of non-trivial repetitive patterns. In the case of neurons, such periodic signals are easily read out and identified by other neurons and can, thereby, be used as code words. Self-similarity of the shrimp areas may simplify the tuning to stay on one code word for slowly changing parameters or to engineer, in a simple way, jumps from one code word to another, enabling in this way simple state-coding. Such a coding is closely related to the coding in terms of Arnol'd tongues for weakly coupled periodic systems (Stoop *et al.*, 2000). There, the coding is easily seen to be invariant with respect to a uniform scaling of the firing frequencies (e.g. by changed driving input applied similarly to all involved neurons), and tongue size and stability is seen to scale with periodicity, which leads to a self-refining Huffman-like efficient code (Huffman, 1952). The particular arrangement of the shrimps in parameter space (Fig. 7) might favor the biological implementation of such a coding scheme.

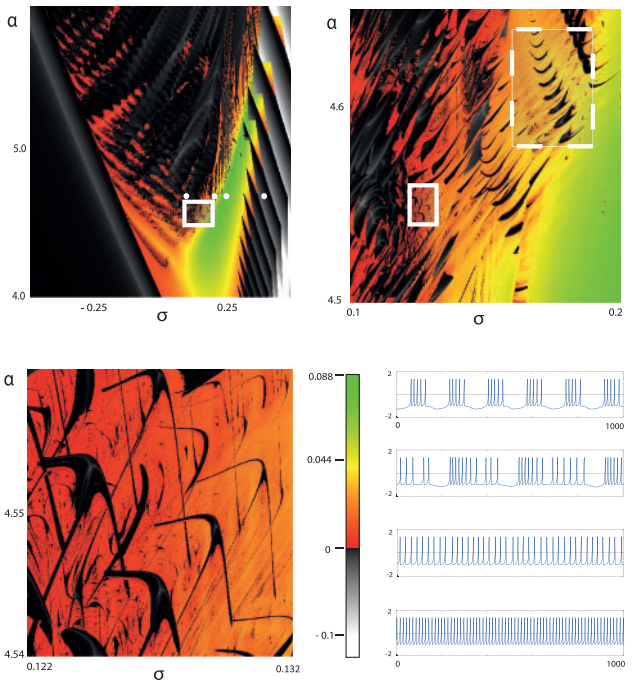

Fig. 6. Largest Lyapunov exponent (Peinke *et al.*, 1992) of Rulkov's neuron model (Rulkov, 2002), three zoom levels (full white boxes). Black color indicates stable periodic systems, ocher unstable systems (generally chaotic). Shrimps-like domains (black) pertain on all levels, where crossing tails reflect non-ergodicity (hysteresis). Spike trains generated at the white dots (left to right corresponding to top to bottom) exhibit the generality of the model

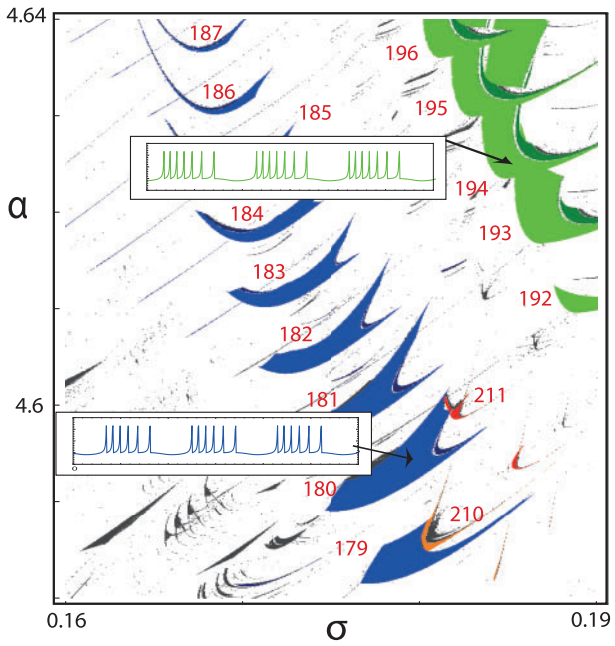

Fig. 7. Periodicity coding for the dashed window of Figure 4. The periodicities follow a period-increasing pathway as known from the Arnol'd tongues, scaling in size with periodicity. Blue and green colors indicate the number of in-burst spikes (six and seven, respectively), red numbers the overall periodicity of the spike train

To relate these observations to bioinformatics, it is important to note that shrimps are of course not restricted to 2D parameter space, they also exist in higher dimensions (Baptista *et al.*, 2003). Moreover, in many applications, objects may not be directly characterized by their fundamental parameters (these are often unknown) but by easily observable features. From parameter coordinates, we arrive at the feature space by means of a feature mapping, which most of the time is implicit. In a not too high-dimensional feature space, feature maps of sufficient smoothness will closely reflect the situation that we have in parameter space [although a formalization of this expectation may require an advanced mathematical framework (Hamilton, 1982)]. A pictorial example is provided by the transformation $(a, b) \rightarrow (a, \mathrm{Log}(1+|b|), ab)$ from 2D into 3D space (Fig. 8).

In what follows, we will exhibit how a severe clustering problem emerges from the shrimps-like parameter domain formed by systems displaying a regular response that most researchers will be unaware of if the data are not compromised by, e.g. wild projection methods. Figure 1 in (Bryan, 2004) may represent such an experimentally observed case. Suppose that we now sample the parameter or feature space with the aim of identifying parameters that lead to a periodic system response. Whatever may be the sampling procedure and the test for periodicity, what will likely result is a situation like Figure 9a: candidate systems will be from primary shrimps or from lesser populated areas hosting smaller shrimps or systems for which the data appear periodic but are actually chaotic (unstable periodic orbits are generically embedded into chaos, and the systems' trajectories can follow such orbits for some time). Taking this situation as a toy example, we now proceed toward the clustering of the data into sets of similar behavior. To this end, we suppose that similar parameters generate more similar behaviors than dissimilar ones. The principle that clustering is thus based on is that the smaller the distance in space (parameter, feature), the more they are coupled and likely to be in the same cluster.

The most commonly used clustering algorithms approach this problem based on this distance or similarity measure alone. As a consequence, they will end up with the 'noisy' data included into the clustering. Clearly, this should be avoided. Let us assume that by a magic 'noise-cleaning' algorithm, we got rid of the noisy part of the data. The interesting observation then is that even in this case, the most prominent clustering algorithms fail in the clustering of convex–concave-bounded sets such as our shrimp-like domains, as they are implicitly based on a linear separability criterion. Although this is evident for the popular k-means algorithm, this also holds for hierarchical agglomerative Wards clustering—irrespective of what distance measure is used. That is, because non-local distances are introduced as soon as one deals with distances between a point and a set, or with distances between sets and sets. In view of the genericity of our situation, such a behavior is detrimental. The naive use of out-of-the-box

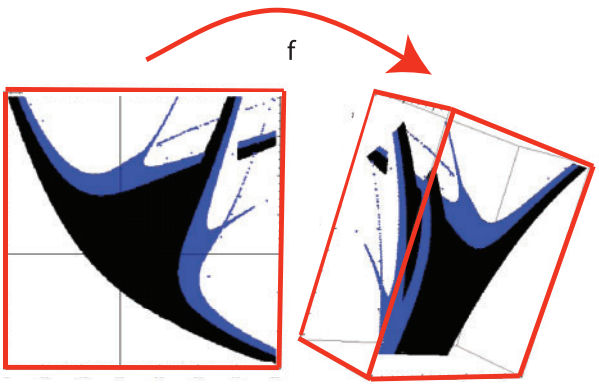

Fig. 8. Artificial feature map example: a 2D shrimp (**a**) is mapped into a 3D feature space shrimp by means of the transformation f: $(a, b) \rightarrow (a, Log(1+|b|), ab)$. Shrimp essentials are preserved under map f; transformations of similar mathematical properties yield comparable results. (Left side: black area: parameters with the same periodicity; blue area: parameters with period-doubled periodicity. Right side: corresponding features.)

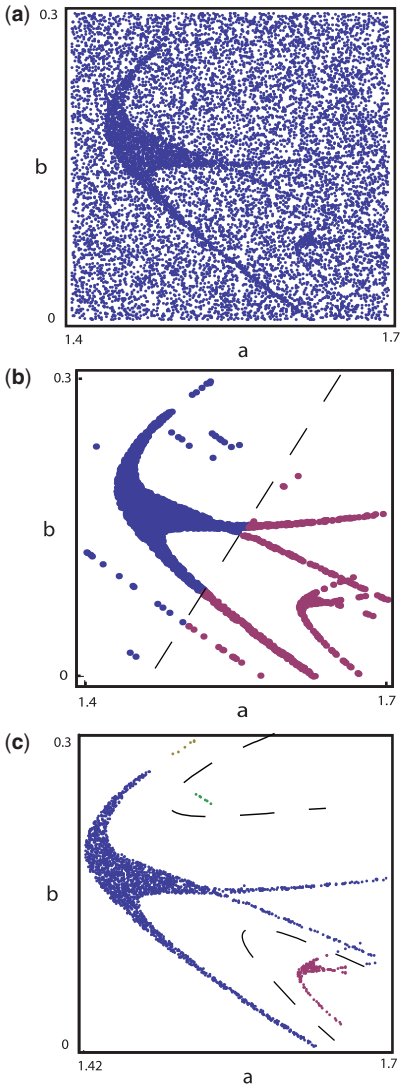

Fig. 9. (**a**) Clustering data: pairs of coordinates $\{a, b\}$. (**b**) Even after discarding the 'noisy' part, clustering via Ward's approach is unsuccessful (virtually independent from the distance measure used). (**c**) Starting from the noisy data of (a), Hebbian learning clustering (see text) reveals the hidden data and provides a proper clustering. Dashed: separation borders between the main clusters (Stoop *et al.*, 2012)

algorithms leads to clustering failure in an important family of objects and obstructs the proper coding interpretation of parameter states.

As we have pointed out above, this geometric situation largely persists in associated feature spaces. For any definition of the distance measure normally used, Wards clustering leads to a failure (see Fig. 9b). Although Ward's clustering appears to be based on local distances, the method becomes intrinsically non-local as soon as the distance from a point to a set has to be evaluated. For an adequate clustering, a clustering algorithm must be used that is entirely based on local neighborhoods and avoids macroscopic notions such as the distance to a set.

There is a solution to both of the problems (i.e. noise and convex–concave boundaries), using, e.g. the following neural networks-inspired algorithm (Landis *et al.*, 2010) that is based on purely local notions of distance. The approach can be separated into four steps. In the first step, from a dataset S with N data items (as shown in Fig. 9a), the pairwise similarities or distances d_{ij} between items or pixels i and j are calculated, using an appropriate distance function d. In our toy case, we use the Euclidean distance, but could also embrace other qualities, for visual data clustering, e.g. pixel color or pixel size. For each item, we then determine for performance reasons the set of k nearest neighbors. By representing each item by a node and connecting each node with its k nearest neighbors, we obtain a basic graph. On this graph, each node's activity is represented as an integrate-and-fire (*I&F*) neuron, and each edge is a symmetric synaptic connection w_{ij} of initial strength $w_{ij} = w_{ji} = \exp((d_{ij}/d_0)^2)$, where d_0 is a constant related to the average network activity. In the second step, *I&F* neuron site dynamics are implemented. The *I&F* neuron is modeled as a resistor-capacitor (RC) circuit that is driven by a current $I = I_{ext} + I_{inner}$. The external input I_{ext} is assumed to be constant and equal for all neurons ('noisy driving'); input I_{inner} describes the input relied over the ('stronger') connections that define the essential topology of the network [c.f. (Stoop *et al.*, 2000)]. At each node, the obtained potential $u(t)$ then follows the equation $\frac{\partial u(t)}{\partial t} = -\frac{u(t)}{RC} + \frac{I_{ext} + \sum_k w_k(t)\delta(t-t^k)}{C}$. t^k are the times of the firing events of the connected neurons. After the firing threshold has been reached, the state of the firing neuron is set to zero, and connected neurons are updated as $u(t') + = w_j R$. Until the next neuron of the population fires, neurons are updated as $u(t') = I_{ext}R(1 - \exp(\frac{-T_k}{RC}) + u(t)\exp(\frac{-T_k}{RC})$, where $t' = t + T_k$ with T_k the time since the last spike. In a third step, a Hebbian-motivated dynamics of the topology of the network is implemented by doubling weights w_{ij} connecting neurons that fire together in a sufficiently small time window (with a cutoff of the process at $w = 1$). This increase of weights is balanced by a weight decay $w(t) = w(0)2^{(-\frac{2\tau t}{\tau_{ext}^2})}$, with τ_{ext} as the firing period of the unconnected neurons. This simple rule finally gives rise to a self-organization and self-amplification mechanism acting on the network, where clusters emerge as sets of strongly connected and synchronous neurons (weights $w_{ij} = 1$). Items that do not have any connections belong to no cluster and are discarded as noise. The method is autonomous—no clustering level or number of clusters to be separated needs to be provided. For more details and for further illustrations, see (Landis *et al.*, 2010). The physics picture underlying this process was recently examined in (Gutiérrez *et al.*, 2011). For the result of such a clustering process, see Figure 9c. In contrast to the application of the classical hierarchical Wards or the k-means clustering leading to result shown in Figure 9b, no noise cleaning preprocessing was necessitated. Spin-motivated clustering systems (e.g. Potts-spin clustering; Blatt *et al.*, 1996; Murua *et al.*, 2008; Ott *et al.*, 2004) work similarly. They start with a ferromagnetic monobloc system that is then heated, upon which the original monobloc splits up into pieces that may then remain unaffected over a considerable temperature interval (and hence are identified as clusters). Upon further heating, they then split up into even smaller clusters and finally into singletons. It is easy to see that the process of heating requires extended computational effort, compared with the neuro-inspired approach.

The question, to what extent the presence of these generic highly interwoven structures of dissimilar behaviors play a role in the parameter inference problem considered in systems biology and elsewhere, is an important and non-trivial one. In a Bayesian context, posterior distributions on parameter space are likely to differ vastly from normality. Therefore, standard inference methods such as the standard Metropolis algorithm could be expected to fail to converge within reasonable time, and that one might have to resort to more sophisticated methods such as genetic population algorithms. In particular, approximate Bayesian computation (ABC) methods (Toni and Stumpf, 2010) that are normally used for Bayesian parameter inference where the stochasticity of the model makes the calculation of the likelihood density prohibitively expensive, might be expected to fail, as the shrimp phenomena are caused by the non-linearities of the systems, and not by stochasticity. Lotka–Volterra systems are often used to demonstrate the efficacy of ABC. These systems are, however, particular in the sense that they come equipped with a distribution of center solutions, a case that is more characteristic for linear than for non-linear systems, where both the parameter and the solution space are more complicated. A natural conclusion that one might draw is that the methods used for combinatorial problem optimization with many local minima (genetic algorithms, particle filters, Monte Carlo methods) have more potential than the ABC methods, and they will also be preferable to Kalman filters or to simple gradient descent estimators (Liu and Niranjan, 2012).

To check these expectations, we performed a survey of applications of ABC methods on our parameter space, where (Toni *et al.*, 2009; Toni and Stumpf, 2010) served as the references of models and methods. Our numerical experiments demonstrate that even in the context of the strongly fractionalized parameter spaces of non-linear systems, the ABC approaches perform well (see the for convenience displayed characteristic results in our Supplemental Material section). This is mainly due to the fact that they are ensemble based. Linear approximation schemes [e.g. singular values decomposition or the independent component analysis (ICA) methods used in source separation of sounds (Kern and Stoop, 2011)] usually performed for dimensional reduction or directly for a gradient descent step, tend to ignore or smear the local structures and are therefore far less suited. Seen in this light, generic data that we presented provide a justification for the superiority of ensemble-based parameter inference methods.

It is mostly in the realm of clustering where the shrimps property of non-linear systems is detrimental in the application of the most prominent and most widely used algorithms.

Funding: This work was supported by the Swiss National Science Foundation (SNF) [grant number 200020-147010/1 to R.S.] that is gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES

Ando,H. *et al.* (2011) Synthetic gene networks as potential flexible parallel logic gates. *Europhys. Lett.*, **93**, 50001.
Baptista,M.S. (1996) Perturbing nonlinear systems, an approach to the control of chaos. PhD Thesis, University of Sao Paulo, Brazil.
Baptista,M.S. *et al.* (2003) Topology of windows in the high-dimensional parameter space of chaotic maps. *Int. J. Bifurcat. Chaos*, **13**, 2681–2688.
Blatt,M. *et al.* (1996) Super-paramagnetic clustering of data. *Phys. Rev. Lett.*, **76**, 3251–3255.
Bonatto,C. *et al.* (2005) Self-similarities in the frequency-amplitude space of a loss-modulated CO2 laser. *Phys. Rev. Lett.*, **95**, 143905.
Bonatto,C. and Gallas,J.A.C. (2008) Periodicity hub and nested spirals in the phase diagram of a simple resistive circuit. *Phys. Rev. Lett.*, **101**, 054101.
Brackley,C.A. *et al.* (2010) Introduction to focus issue: dynamics in systems biology. *Chaos*, **20**, 045101.
Bryan,J. (2004) Problems in gene clustering based on gene expression data. *J. Multivar. Anal.*, **90**, 44–66.
Cardoso,J. *et al.* (2009) Complex periodic structures in bi-dimensional bifurcation diagrams of a RLC circuit model with a nonlinear NDC device. *Phys. Lett. A*, **373**, 2050–2053.
Decroly,O. and Goldbeter,A. (1982) Birhythmicity, chaos, and other patterns of temporal selforganization in a multiply regulated biochemical system. *Proc. Natl Acad. Sci. USA*, **79**, 6917–6921.
De la Fuente,I.M. (1999) Diversity of temporal self-organized behaviors in a biochemical system. *Biosystems*, **50**, 83–97.
Ellner,S.P. and Guckenheimer,J. (2006) *Dynamic Models in Biology*. Princeton University Press, Princeton, NJ.
Erzberger,A. *et al.* (2013) Genetic redundancy strengthens the circadian clock leading to a narrow entrainment range. *J. R. Soc. Interface*, **10**, 20130221.
Furusawa,C. and Kaneko,K. (2012) A dynamical-systems view of stem cell biology. *Science*, **338**, 215–217.
Gallas,J.A.C. (1993) structure of the parameter space of the Hénon map. *Phys. Rev. Lett.*, **70**, 2714–2717.
Gallas,J.A.C. (1995) Structure of the parameter space of a ring cavity. *Appl. Phys. B*, **60**, 203.
Gallas,J.A.C. (2010) The structure of infinite periodic and chaotic hub cascades in phase diagrams of simple autonomous flows. *Int. J. Bifurcat. Chaos*, **20**, 197.
Gaspard,P. *et al.* (1984) Bifurcation phenomena near homoclinic systems: a two-parameter analysis. *J. Stat. Phys.*, **35**, 697–727.
Granada,A.E. *et al.* (2011) Circadian desynchronization. *Interface Focus*, **1**, 153–166.
Gutiérrez,R. *et al.* (2011) Emerging meso- and macroscales from synchronization of adaptive networks. *Phys. Rev. Lett.*, **107**, 234103.
Hamilton,R.S. (1982) The inverse function theorem of Nash and Moser. *Bull. Am. Math. Soc.*, **7**, 65–222.
Hénon,M. (1976) A two-dimensional mapping with a strange attractor. *Commun. Math. Phys.*, **50**, 69–77.
Huffman,D.A. (1952) A method for the construction of minimum-redundancy codes. *Proc. IRE*, **40**, 1098–1101.
Kern,A. and Stoop,R. (2011) Principles and typical computational limitations of sparse speaker separation based on deterministic speech features. *Neural Comput.*, **23**, 2358–2389.
Konopka,R.J. and Benzer,S. (1971) Clock mutants of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **68**, 2112–2116.
Kuznetsov,A. (2004) *Elements of Applied Bifurcation Theory*. Springer, Berlin.
Landis,F. *et al.* (2010) Hebbian self-organizing integrate-and-fire networks. *Neural Comput.*, **22**, 273–288.
Liu,X. and Niranjan,M. (2012) State and parameter estimation of the heat shock response system using Kalman and particle filters. *Bioinformatics*, **28**, 1501–1507.
Maranhao,D.M. *et al.* (2008) Experimental observation of a complex periodic window. *Phys. Rev. E*, **77**, 037202.
Markus,M. *et al.* (1985) Properties of strange attractors in yeast glycolysis. *Biophys. Chem.*, **22**, 95–105.
Martignoli,S. *et al.* (2013) Pitch sensation involves stochastic resonance. *Sci. Rep.*, **3**, 2676.
Martignoli,S. and Stoop,R. (2008) Phase-locking and Arnold coding in prototypical network topologies. *Discrete Continuous Dyn. Syst. B*, **9**, 145–162.
Murua,A. *et al.* (2008) On Potts model clustering, kernel K-means, and density estimation. *J. Comput. Graph. Stat.*, **17**, 629–658.
Nishio,Y. *et al.* (1990) Rigorous analysis of windows in a symmetric circuit. *IEEE Trans. Circuits Syst.*, **37**, 473–487.
Nowotny,T. *et al.* (2005) Self-organization in the olfactory system: one shot odor recognition in insects. *Biol. Cybern.*, **93**, 436.
Ott,T. *et al.* (2004) Sequential superparamagnetic clustering for unbiased classification of high-dimensional chemical data. *J. Chem. Inf. Comput. Sci.*, **44**, 1358–1364.
Peinke,J. *et al.* (1992) *Encounter with Chaos: Self-Organized Hierarchical Complexity in Semiconductor Experiments*. Springer, Berlin.
Raj,A. and van Oudenaarden,A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**, 216–226.
Romano,M.C. *et al.* (2009) Queueing phase transition: theory of translation. *Phys. Rev. Lett.*, **102**, 198104.
Rulkov,N.F. (2002) Modeling of spiking-bursting neural behavior using two-dimensional map. *Phys. Rev. E*, **65**, 041922.
Rulkov,N.F. *et al.* (2004) Oscillations in large-scale cortical networks: map-based model. *J. Comput. Neurosci.*, **17**, 203.
Shilnikov,A. and Rulkov,N.F. (2003) Origin of chaos in a two-dimensional map modeling spiking-bursting neural activity. *Int. J. Bifurcat. Chaos*, **13**, 3325–3340.
Shilnikov,L.P. (1965) A case of the existence of a denumerable set of periodic motions. *Sov. Math. Dokl.*, **6**, 163.
Shilnikov,L.P. (1967) The existence of a denumerable set of periodic motions in four-dimensional space in an extended neighborhood of a saddle-focus. *Sov. Math. Dokl.*, **8**, 54.
Shilnikov,L.P. *et al.* (1998) *Methods of Qualitative Theory in Nonlinear Dynamics I*. World Scientific, Singapore.
Shilnikov,L.P. *et al.* (2001) *Methods of Qualitative Theory in Nonlinear Dynamics II*. World Scientific, Singapore.
Silva,A.P.D. and McClenaghan,N.D. (2004) Molecular-scale logic gates. *Chemistry*, **10**, 574–586.
Spiller,D.G. *et al.* (2010) Measurement of single-cell dynamics. *Nature*, **465**, 736–745.
Stoop,R. *et al.* (2000a) Generic origins of irregular spiking in neocortical networks. *Biol. Cybern.*, **83**, 481–489.
Stoop,R. *et al.* (2000b) Noise-driven neocortical interaction: a simple generation mechanism for complex neuron spiking. *Acta. Biotheor.*, **48**, 149–171.
Stoop,R. *et al.* (2010) Real-world existence and origins of the spiral organization of shrimp-shaped domains. *Phys. Rev. Lett.*, **105**, e074102.
Stoop,R. *et al.* (2012) Shrimps: occurrence, scaling and relevance. *Int. J. Bifurcat. Chaos*, **22**, 1230032.
Suter,D.M. *et al.* (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**, 472–474.
Tainaka,G. *et al.* (2006) Synchronization and propagation of bursts in networks of coupled map neurons. *Chaos*, **16**, 013113.
Tei,H. *et al.* (1997) Circadian oscillation of a mammalian homologue of the Drosophila period gene. *Nature*, **389**, 512–516.
Toni,T. and Stumpf,M.P.H. (2010) Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, **26**, 104–110.
Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.
Ueda,H.R. *et al.* (2005) System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat. Genet.*, **37**, 187–192.
Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.